

Imputation results, visualization of RNA-Seq data

Natalja Kurbatova PhD
EBI, Cambridge, UK

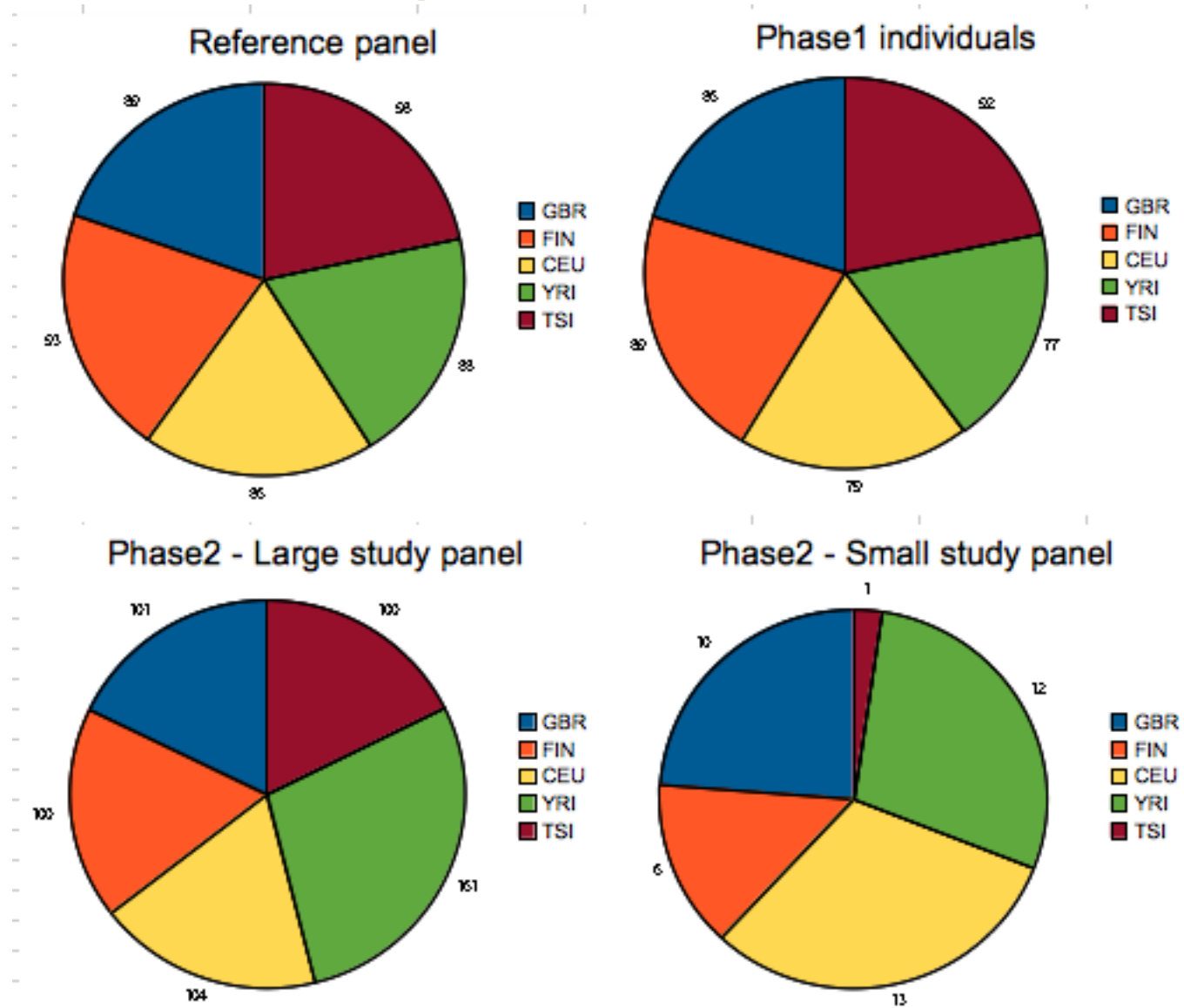
Barcelona, Geuvadis meeting, July 2012



Imputation

- Study panel: Omni2.5 genotypes data for 1000g phase2 individuals (phased and unphased)
- Reference panel: integrated haplotypes, SNPs, indels and SV 1000g phase1 individuals
- Impute2 software
- Small study panel (42 individuals) gave us very poor quality of imputation (32% of imputed SNPs $INF > 0$)
- Increasing of the study panel (1856 individuals) improved the quality of imputation (84% of imputed SNPs $INF > 0$)
- Imputation procedure and location of the files can be found on wiki
-

Panels used in imputation



Overview of imputation using IMPUTE2



This figure over-simplifies what IMPUTE2 does. The output for each genotype is actually a probability distribution:

Genotype	0	1	2
Probability	0.01	0.18	0.81

This captures the uncertainty in the prediction.

Impute2 software

Several important points emerge from this description. First, the accuracy with which the study haplotypes are phased at SNPs in T should determine how well they can be matched to haplotypes in the reference panel, which should in turn influence the accuracy of imputation at SNPs in U . Second, accounting for the unknown phase of the SNPs in T can be computationally expensive; if the haplotypes at these SNPs were known, most methods would be able to impute genotypes at SNPs in U more quickly. Third, many existing methods do not use all of the available information to phase the study genotypes at SNPs in T . In principle, a phasing algorithm should be able to “learn” about desirable phasing configurations for a given study individual by pooling information across the reference panel and all other individuals in the study, and the phasing accuracy should increase with the sample size; in standard practice, most imputation methods gain phasing information about each study individual only from the reference panel, and phasing accuracy does not depend on the size of the study sample. (This description applies to imputation methods based on hidden Markov models, or “HMMs” [6],[11]; non-HMM methods often discard other kinds of information.) The BEAGLE imputation model [12],[13] is one notable exception to this point, and we discuss its alternative modeling strategy in detail in this work.

<http://www.plosgenetics.org/article/info%3Adoi%2F10.1371%2Fjournal.pgen.1000529>

Vcf files: merged phase1 423 individuals plus 42 phase2 individuals

```
##ALT=<ID=DEL,Description="Deletion relative to the reference">
##INFO=<ID=EAF,Number=.,Type=Float,Description="Expected Allele Frequency from Impute2">
##INFO=<ID=IMP,Number=.,Type=Integer,Description="Imputed for phase2 (1) or measured (0)">
##INFO=<ID=CERTAINTY,Number=.,Type=Float,Description="Average certainty of best-guess genotypes from Impute2">
##INFO=<ID=INF,Number=.,Type=Float,Description="Measure of the observed statistical information associated with the allele frequency estimate from Impute2">
##INFO=<ID=ADD_INFO,Number=A,Type=String,Description="Additional info">

##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=PP,Number=.,Type=String,Description="Posterior probabilities">
##FORMAT=<ID=BD,Number=1,Type=Float,Description="Genotype dosage from beagle">
```

Vcf files: merged phase1 423 individuals plus 42 phase2 individuals

```
##ALT=<ID=DEL,Description="Deletion relative to the reference">
##INFO=<ID=SF,Number=.,Type=String,Description="Source File (index to sourceFiles, f when filtered)">
##INFO=<ID=AC,Number=.,Type=Integer,Description="Allele count in genotypes">
##INFO=<ID=AN,Number=1,Type=Integer,Description="Total number of alleles in called genotypes">
##INFO=<ID=LDAF,Number=1,Type=Float,Description="MLE Allele Frequency Accounting for LD">
...
##INFO=<ID=HOMLEN,Number=.,Type=Integer,Description="Length of base pair identical micro-homology at
event breakpoints">
##INFO=<ID=HOMSEQ,Number=.,Type=String,Description="Sequence of base pair identical micro-homology
at event breakpoints">
##INFO=<ID=SVLEN,Number=1,Type=Integer,Description="Difference in length between REF and ALT
alleles">
##INFO=<ID=SVTYPE,Number=1,Type=String,Description="Type of structural variant">
##INFO=<ID=AA,Number=1,Type=String,Description="Ancestral Allele,
ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/pilot_data/technical/reference/ance
stral_alignments/README">
##INFO=<ID=AF,Number=1,Type=Float,Description="Global Allele Frequency based on AC/AN">
##INFO=<ID=EUR_AF,Number=1,Type=Float,Description="Allele Frequency for samples from EUR based on
AC/AN">
...
##FORMAT=<ID=GT,Number=1,Type=String,Description="Genotype">
##FORMAT=<ID=DS,Number=1,Type=Float,Description="Genotype dosage from MaCH/Thunder">
##FORMAT=<ID=GL,Number=.,Type=Float,Description="Genotype Likelihoods">
```


Vcf files: merged phase1 423 individuals plus 42 phase2 individuals

#CHROM	POS	ID	REF	ALT	QUAL	FILTER	INFO
1	10611	rs189107123	C	G	100	PASS	AA=.;AC=15;AF

AA=.;AC=15;AF=0.02;AFR_AF=0.01;AMR_AF=0.03;AN=930;ASN_AF=0.01;AVGPOST=0.9330;
ERATE=0.0048;EUR_AF=0.02;LDAF=0.0479;RSQ=0.3475;SF=0,1;SNPSOURCE=LOWCOV;THETA=0.0077;VT=SNP;
CERTAINTY=0.987;EAF=0.009;IMP=1;INF=0.385;

Phase 1 individuals

FORMAT	HG00096	HG00097
GT:GL:DS:PP:BD	0 0:-0.48,-0.48,-0.48:0.050	0 1:-0.24,-0.44,-1.16:0.750

Phase 2 individuals

FORMAT	HG00105	HG00107
GT:GL:DS:PP:BD	0 0:.:.:0.779,0.219,0.002:0.223	0 1:.:.:0.053,0.935,0.012:0.959

Visualisation of bam/wiggle files

<http://www.ebi.ac.uk/~natalja/Geuvadis.html>

The main question: bam files or wiggle files?